

The invention in which an exclusive right is claimed is defined by the following:

1. A method for determining one or more performance metrics for a distributed application in which distributed application data are transferred from a first site to a second site over a network, comprising the steps of:

(a) enabling a user to transmit a request for the distributed application data desired by the user, said request being transmitted from the second site to the first site over the network;

(b) in response to the request, transmitting the distributed application data from the first site to the second site over the network, if the distributed application data are not already accessible at the second site;

(c) including machine instructions that define a performance monitoring function with the distributed application data that were requested and transmitted over the network to the second site; and

(d) executing the machine instructions at the second site, to implement the performance monitoring function and to determine the one or more performance metrics for the distributed application without using the performance monitoring function to request any distributed application data from any other site.

2. The method of Claim 1, wherein the performance monitoring function at the second site is initiated after the distributed application data are accessed at the second site.

3. The method of Claim 1, further comprising the step of collecting the one or more performance metrics for the distributed application over the network.

4. The method of Claim 1, wherein the step of collecting the performance metrics includes the step of applying a probabilistic sampling parameter to determine whether performance metrics are collected from each of a plurality of sites.

5. The method of Claim 4, wherein the probabilistic sampling parameter is applied on a per-session basis.

6. The method of Claim 4, wherein the probabilistic sampling parameter is applied on a per-request basis.

7. The method of Claim 1, wherein the performance monitoring function at the second site determines one or more of:

- (a) a fetch latency, corresponding to a time period required to fetch the distributed application data from the first site over the network;
- (b) a render latency, corresponding to a time period to process previously fetched distributed application data and any additional components required for rendering and display;
- (c) a dwell latency, corresponding to a time period exhibited by a user requesting the distributed application data, before requesting other distributed application data;
- (d) a per-image fetch latency, corresponding to a time period for fetching a specific image referenced in the distributed application data;
- (e) an image arrival time, corresponding to a time at which a specific image, loaded as a part of accessing the distributed application data, arrives at the second site;
- (f) a navigation status, corresponding to an event that brought a user to the distributed application data;
- (g) a cache status, corresponding to a determination of whether the distributed application data was cached at the second site;
- (h) a window resize event, corresponding to a determination of whether the user resized a window in which the distributed application data are accessed;
- (i) a page stop event, corresponding to a determination of whether the user aborted loading the distributed application data;
- (j) an image event, corresponding to one of a determination of whether an error occurred while loading an image referenced in the distributed application data, and a determination of whether loading of an image was aborted; and
- (k) a JavaScript error event, corresponding to a determination of whether an error occurred during interpretation of JavaScript included in the distributed application data.

8. The method of Claim 3, further comprising the step of determining whether to collect a performance metric from the second site as a function of a specific performance metric that was determined at the second site.

9. The method of Claim 1, further comprising the step of determining whether the distributed application data are already cached at the second site or must be transferred from the first site, before determining a performance metric.

10. The method of Claim 1, wherein the distributed application data have a markup language format.

11. The method of Claim 1, further comprising the steps of determining a performance metric at the first site; and combining the performance metric determined at the second site with a performance metric determined at the first site to determine a correlated performance metric.

12. The method of Claim 1, wherein said one or more performance metrics is determined substantially without any apparent affect on the access of the distributed application data at the second site.

13. A machine-readable medium on which are stored machine instructions for inclusion with distributed application data that are transferred from one site to another, said machine instructions causing:

(a) a performance monitoring function to be implemented when the distributed application data are accessed; and

(b) the performance monitoring function to determine one or more performance metrics for a distributed application in which the distributed application data are transferred between sites and accessed at one of the sites, without using the performance monitoring function to request any distributed application data from any other site.

14. A system for determining one or more performance metrics for a distributed application in which distributed application data are transferred from a first site to a second site over a network, comprising:

- (a) a memory;
- (b) a network interface; and
- (c) a processing device that is coupled to the memory, and the

network interface, said network interface being adapted to enable communication over the network, wherein at the second site, the processing device causes a request for the distributed application data to be transmitted over the network through the network interface to the first site, said processing device at the first site responding by transmitting the distributed application data along with machine instructions that cause the processing device at the second site to perform a performance monitoring function when executed by said processing device as the distributed application data are accessed at the second site, said performance monitoring function determining said at least one performance metric and being implemented without requiring any affirmative action by a user of the processing device and without using the performance monitoring function to request any distributed application data from any other site.

15. The system of Claim 14, wherein the machine instructions cause the processing device at the second site to transmit said at least one performance metric over the network to a data center serving as a collection site for performance metrics, said data center comprising one of the first site and a separate site that is tasked with collecting the performance metrics.

16. The system of Claim 15, wherein a probabilistic sampling parameter is applied to determine whether the performance metric is collected at the data center.

17. The system of Claim 16, wherein the probabilistic sampling parameter is applied on a per-session basis.

18. The system of Claim 16, wherein the probabilistic sampling parameter is applied on a per-request basis.

19. The system of Claim 14, wherein the machine instructions executed by the processing device at the second site cause a determination of one or more of:

- (a) a fetch latency, corresponding to a time period required to fetch the distributed application data from the first site over the network;
- (b) a render latency, corresponding to a time period to process previously fetched distributed application data and any additional components required for rendering and display;
- (c) a dwell latency, corresponding to a time period exhibited by a user requesting the distributed application data, before requesting other distributed application data;
- (d) a per-image fetch latency, corresponding to a time period for fetching a specific image referenced in the distributed application data;
- (e) an image arrival time, corresponding to a time at which a specific image, loaded as a part of accessing the distributed application data, arrives at the second site;
- (f) an image event, corresponding to at least one of a failure of an image to be returned in response to a request, and a user aborting loading of an image;
- (g) a navigation status, corresponding to an event that brought a user to the distributed application data;
- (h) a cache status, corresponding to a determination of whether the distributed application data was already cached in the memory at the second site;
- (i) a window resize event, corresponding to a determination of whether the user resized a window in which the distributed application data accessed are rendered on the display at the second site;
- (j) a page stop event, corresponding to a determination of whether the user aborted loading the distributed application data from the first site;
- (k) an image error event, corresponding to a determination of whether an error occurred while loading an image referenced in the distributed application data; and

(l) a JavaScript error event, corresponding to a determination of whether an error occurred during interpretation of JavaScript included in the distributed application data.

20. The system of Claim 14, wherein the machine instructions cause the processing device at the second site to determine whether the distributed application data are cached at the second site or must be transferred from the first site, before determining said one or more performance metrics.

21. A method for determining and collecting at least one performance metric related to access of a Web page by a browser program on a client device, including at least one of a compound performance metric and a correlated performance for a network, comprising the steps of:

(a) enabling a user to request transfer of the Web page from a server device to the client device over a network;

(b) including machine instructions with the Web page when the Web page is transferred to the client device;

(c) when the Web page is loaded by the client device for rendering by the browser program, causing the client device to execute the machine instructions to carryout a browser monitoring function, said browser monitoring function being implemented without requiring any affirmative action by a user of the client device;

(d) determining said at least one performance metric on the client device with the browser monitoring function without using the browser monitoring function to request any Web page from any other site; and

(e) if a correlated performance metric is to be determined:

(i) determining a server performance metric on a server; and

(ii) combining the server performance metric with said at least one performance metric to determine the correlated performance metric.

22. The method of Claim 21, further comprising the step of transmitting said at least one performance metric from the client device to a remote site over the network.

23. The method of Claim 22, wherein the remote site comprises a data center, further comprising the step of analyzing said at least one performance metric to determine performance data for the Web page, including the correlated performance metric.

24. The method of Claim 22, further comprising the step of enabling a determination to be made of whether said at least one performance metric will be transmitted to the data center for processing, based upon a probabilistic sampling parameter.

25. The method of Claim 24, wherein the probabilistic sampling parameter is applied on a per-session basis to determine if said at least one performance metric will be transmitted to the data center.

26. The method of Claim 24, wherein the probabilistic sampling parameter is applied on a per-request basis to determine if said at least one performance metric will be transmitted to the data center.

27. The method of Claim 22, wherein a plurality of different kinds of performance metrics can be determined by the browser monitoring function, further comprising the step of enabling the data center to selectively accept a performance metric as a function of the kind of performance metric being transmitted to the data center.

28. The method of Claim 21, wherein the step of determining said at least one performance metric is done without the client device providing any indication to the user of the client device that said at least one performance metric is being determined.

29. The method of Claim 21, wherein when determining said at least one performance metric, the client device determines one or more of:

(a) a fetch latency, corresponding to a time period required to fetch a base Web page document from a server over the network;

(b) a render latency, corresponding to a time period required to fetch and display all contents referenced within an Hypertext Markup Language (HTML) document on the client device;

(c) a dwell latency, corresponding to a time period exhibited by the user viewing the Web page, before navigating to a different Web page with the browser program;

(d) a per-image fetch latency, corresponding to a time period for fetching a specific image referenced in the Web page;

(e) an image arrival time, corresponding to a time at which a specific image, loaded as a part of rendering the Web page, arrives on the browser;

(f) a navigation status, corresponding to an event that brought the user to the Web page;

(g) a cache status, corresponding to a determination of whether the Web page was cached by the browser program or by a proxy;

(h) a window resize event, corresponding to a determination of whether the user resized a window in which the Web page is rendered;

(i) a page stop event, corresponding to a determination of whether the user aborted loading of the Web page;

(j) an image event, corresponding to one of a determination of whether an error occurred while loading an image referenced in the Web page, and a determination of whether loading of an image was aborted; and

(k) a JavaScript error event, corresponding to a determination of whether an error occurred during interpretation of JavaScript included in the Web page.



30. The method of Claim 21, further comprising the step of determining whether the Web page was previously cached by the client device.

31. The method of Claim 21, wherein said at least one performance metric comprises a performance metric for each image included in the Web page.

32. The method of Claim 21, further comprising the steps of:

(a) including a monitor cookie with the Web page that is transferred to the client device from the server device and indicates that the Web page is a monitored document;

(b) detecting the monitor cookie when the Web page is transferred to the client device; and

(c) causing the browser monitor function to determine that said at least one performance metric is to be determined for the Web page in response to the monitor cookie being detected.

33. The method of Claim 21, further comprising the steps of:

(a) executing a server monitoring function on a server device that is transferring the Web page to the client device;

(b) determining the server performance metric related to the transfer of the Web page to the client device from the server device with the server monitoring function; and

(c) combining said server performance metric with said at least one performance metric determined by the browser monitoring function on the client device, to determine the correlated performance of the network.

34. The method of Claim 21, wherein the step of combining the server performance metric determined by the server monitoring function with said at least one performance metric determined by the browser monitoring function determines a network latency.

35. A memory medium on which are stored machine readable instructions, which when executed by a client computing device, cause the client computing device to carryout a browser monitoring function, said browser monitoring function being implemented without requiring any affirmative action by a user of the client computing device and being used for determining at least one performance metric on the client computing device with the browser monitoring function, said at least one performance metric being related to access of a distributed application by a browser program executed on the client computing device and enabling at least one of a compound performance metric and a correlated performance metric be determined without using the browser monitoring function to request any distributed application from any other site.

36. The memory medium of Claim 35, wherein the machine readable instructions cause said at least one performance metric to be transmitted to a remote site over a network for determination of the correlated performance metric.

37. The memory medium of Claim 35, wherein said at least one performance metric is determined without the client device providing any indication to a user of the client device that said at least one performance metric is being determined.

38. The memory medium of Claim 35, wherein the machine readable instructions determine one or more of the following performance metrics:

- (a) a fetch latency, corresponding to a time period required to fetch a base distributed application document from a server over the network;
- (b) a render latency, corresponding to a time period required to fetch and display all contents referenced within an HTML document on the client device;
- (c) a dwell latency, corresponding to a time period exhibited by the user viewing the distributed application, before navigating to a different distributed application with the browser program;

(d) a per-image fetch latency, corresponding to a time period for fetching a specific image referenced in the distributed application;

(e) an image arrival time, corresponding to a time at which a specific image, loaded as a part of rendering the distributed application, arrives on the browser;

(f) a navigation status, corresponding to an event that brought the user to the distributed application;

(g) a cache status, corresponding to a determination of whether the distributed application was cached by the browser program or by a proxy;

(h) a window resize event, corresponding to a determination of whether the user resized a window in which the distributed application is rendered;

(i) a page stop event, corresponding to a determination of whether the user aborted loading of the distributed application;

(j) an image event, corresponding to one of a determination of whether an error occurred while loading an image referenced in the distributed application data, and a determination of whether loading of an image was aborted; and

(k) a JavaScript error event, corresponding to a determination of whether an error occurred during interpretation of JavaScript included in the distributed application.

39. The memory medium of Claim 35, wherein the machine readable instructions cause the client computing device to determine whether the distributed application was previously cached by the client computing device.

40. The memory medium of Claim 35, wherein said at least one performance metric includes a performance metric for each image in the distributed application.

41. The memory medium of Claim 35, wherein the machine readable instructions cause the client computing device to:

(a) detect whether a monitor cookie is included with the distributed application that is transferred to the client computing device, said monitor cookie indicating that the distributed application is a monitored document; and

(b) cause the browser monitor function to determine that said at least one performance metric is to be determined for the distributed application in response to the monitor cookie being detected.

42. A system for determining and collecting at least one performance metric related to access of a Web page by a browser program, comprising:

(a) a memory;

(b) a network interface; and

(c) a processing device that is coupled to the memory, and the network interface, said network interface being adapted to couple to a remote storage at a server to retrieve the Web page, said Web page including machine instructions that perform a browser monitoring function and which are executed by the processing device when the Web page is loaded by the processing device for rendering, said browser monitoring function determining said at least one performance metric and being implemented without requiring any affirmative action by a user of the processing device and without using the browser monitoring function to request any Web page from any other site, said at least one performance metric including at least one of compound performance metric and a correlated performance metric.

43. The system of Claim 42, wherein the machine instructions further cause the processing device to transmit said at least one performance metric from the processing device to a remote site over a network through the network interface.

44. The system of Claim 43, further comprising a computing device disposed remotely at a data center, said computing device receiving and analyzing said at least one performance metric to determine performance data for the Web page, said performance data including the correlated performance metric for the network.

45. The system of Claim 44, wherein a determination of whether said at least one performance metric will be accepted for processing by the data center is based upon a probabilistic sampling parameter, ensuring that performance metrics transmitted to the data center are randomly sampled.

46. The system of Claim 45, wherein the probabilistic sampling parameter is applied on a per-user basis to determine if said at least one performance metric is accepted for processing by the data center.

47. The system of Claim 45, wherein the probabilistic sampling parameter is applied on a per-request basis to determine if said at least one performance metric will be accepted for processing by the data center.

48. The system of Claim 44, wherein a plurality of different kinds of performance metrics can be determined by the browser monitoring function, and wherein the data center selectively accepts said at least one performance metric, based upon a specific kind of performance metric that is being transmitted to it for processing.

49. The system of Claim 42, wherein said at least one performance metric is determined by the processing device without providing any indication to a user of the processing device that said at least one performance metric is being determined.

50. The system of Claim 42, wherein said at least one performance metric includes one or more of:

- (a) a fetch latency, corresponding to a time period required to fetch a base Web page document over the network;
- (b) a render latency, corresponding to a time period to process previously fetched contents of the Web page and any additional components required for rendering and display;
- (c) a dwell latency, corresponding to a time period exhibited by a user viewing the Web page, before navigating to a different Web page;
- (d) a per-image fetch latency, corresponding to a time period for fetching a specific image referenced in the Web page;
- (e) an image arrival time, corresponding to a time at which a specific image, loaded as a part of rendering the Web page, arrives for rendering;
- (f) a navigation status, corresponding to an event that brought a user to the Web page;
- (g) a cache status, corresponding to a determination of whether the Web page was cached in the memory by a browser program or by a proxy;
- (h) a window resize event, corresponding to a determination of whether a user resized a window in which the Web page is rendered;
- (i) a page stop event, corresponding to a determination of whether a user aborted loading of the Web page;
- (j) an image event, corresponding to one of a determination of whether an error occurred while loading an image referenced in the Web page, and a determination of whether loading of an image was aborted; and
- (k) a JavaScript error event, corresponding to a determination of whether an error occurred during interpretation of JavaScript included in the Web page.

51. The system of Claim 42, wherein the machine instructions further cause the processing device to determine if the Web page was previously cached in the memory by the processing device, before determining said at least one performance metric.

52. The system of Claim 42, wherein said at least one performance metric comprises a performance metric for an object included in the Web page.

53. The system of Claim 42, wherein the machine instructions further cause the processing device to:

(a) detect whether a monitor cookie is included with the Web page, said monitor cookie indicating that the Web page is a monitored document; and

(b) cause the processing device to determine that said at least one performance metric is to be determined for the Web page in response to the monitor cookie being detected.

54. The system of Claim 44, further comprising:

(a) a server computing device that is remote from the processing device and coupled in communication with the processing device and with the data center over a network through the network interface, said server computing device executing a server monitoring function in regard to transferring the Web page to the processing device over the network;

(b) said server computing device determining a server performance metric related to the transfer of the Web page to the processing device from the server computing device; and

(c) said server computing device transmitting said server performance metric to the data center site for processing.

55. The system of Claim 54, wherein the data center combines a performance metric determined by the browser monitoring function executed by the processing device with the server performance metric determined by the server computing function to determine the correlated performance metric.

56. The system of Claim 54, further comprising a caching proxy disposed between the server computing device and the processing device, said caching proxy executing a caching proxy monitoring function that determines at least one performance metric related to a performance of the caching proxy.

57. A method for automatically directing a computing device that is about to download distributed application data so that the computing device obtains the distributed application data from one of a plurality of sources nominally providing the distributed application data, comprising the steps of:

(a) determining performance metrics relating to downloading other distributed application data from each of at least two of the plurality of sources by the computing device; and

(b) as a function of the performance metrics for said at least two of the plurality of sources, automatically directing the computing device to request download of the distributed application data from a preferred source having a better performance metric compared to other of the plurality of sources.

58. The method of Claim 57, further comprising the steps of:

(a) conveying machine instructions to the computing device with earlier distributed application data downloaded by the computing device;

(b) executing the machine instructions on the computing device to determine the performance metrics when other distributed application data are downloaded from said at least two of the plurality of sources; and

(c) executing machine instructions included with the other distributed application data that cause the distributed application data to be downloaded from the preferred source.

59. The method of Claim 58, wherein the machine instructions define a performance-based redirection browser monitor and wherein the distributed application data and the other distributed application data define Web pages.

60. A memory medium on which the machine instructions of steps (a) and (b) of Claim 58 are stored.



61. A system for automatically directing distributed application data to be downloaded from one of a plurality of sources providing the distributed application data, so as to achieve a relatively better performance in downloading the distributed application data, comprising:

- (a) a monitor on which are displayed text and graphics;
- (b) a memory in which are stored machine instructions;
- (c) a network interface that couples to a network over which the distributed application data are conveyed; and

- (d) a processor coupled to the monitor, the network interface, and the memory, said processor carrying out a plurality of functions, including:

- (i) determining performance metrics relating to downloading other distributed application data from each of at least two of a plurality of sources; and

- (ii) as a function of the performance metrics for said at least two of the plurality of sources, automatically directing a request to download the distributed application data from a preferred source having a better performance metric than any other of the plurality of sources.